Report: Optimising NYC Taxi Operations Shilpa Deshmukh

Include your visualisations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

## Data Preparation

* 1. Loading the dataset

**Sample the data and combine the files**  
We first read all the parquest file from the folder os.chdir(



## Data Cleaning

### Fixing Columns

* + 1. **Fix the index**  
       remove the cplumn unnamed:0 to fix the index
    2. **Combine the two airport\_fee columns**

**To combine the airport\_fee column check if the values in both column is same if yes keep the value as it is if not than add noth the values and put it in one single coloumn. And delete the other where we haven’t done any work.**

### Handling Missing Values

* + 1. **Find the proportion of missing values in each column**

* + 1. **Handling missing values in passenger\_count**

* + 1. **Handle missing values in RatecodeID**
    2. **Impute NaN in congestion\_surcharge**

### Handling Outliers and Standardising Values

* + 1. **Check outliers in payment type, trip distance and tip amount columns**  
       A graph with a line

       AI-generated content may be incorrect.

A graph of a box plot

AI-generated content may be incorrect.

A graph of a box plot

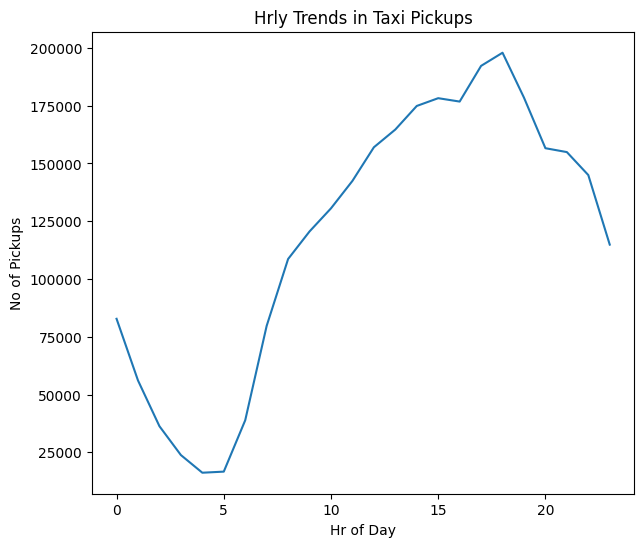
AI-generated content may be incorrect.

## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical**
    2. **Analyse the distribution of taxi pickups by hours, days of the week, and months**

The no. of pickups are highest between 16-18hrs and lowest between 4-5hrs.



**The no. of Pickups are more on the thrusday and almost similar trend goes for Wednesday and Friday . Tuesday and satursay , Monday and Sunday.** A graph of blue bars

AI-generated content may be incorrect.

**The Month of October and May see the highest no. of pickups and the month July, August and September see the lower pickup number.**

A graph of blue bars

AI-generated content may be incorrect.

* + 1. **Filter out the zero/negative values in fares, distance and tips**

**There are zero values in fares distance and tips columns as below . we find the zero values by using the below code**

zero\_<<col name>> = (df[col name'] == 0).sum()

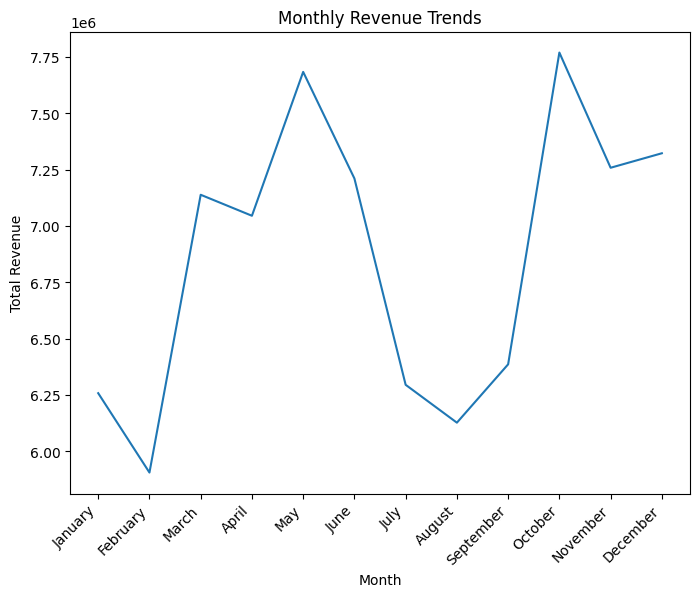
**the Negative value is calculated using the below code**

negative\_<<col name>>= (df[col name'] < 0).sum()

**the col name is the name of the column where we wont to get the zero or negative values.**

**The below code will filter out the non zero and non negative values and there wont be any zero or negative vaues in the dataset now.**

final\_df\_without\_zero = df.loc[(df['fare\_amount'] != 0) & (df['tip\_amount'] != 0) & (df['total\_amount'] != 0) & (df['trip\_distance'] != 0)]

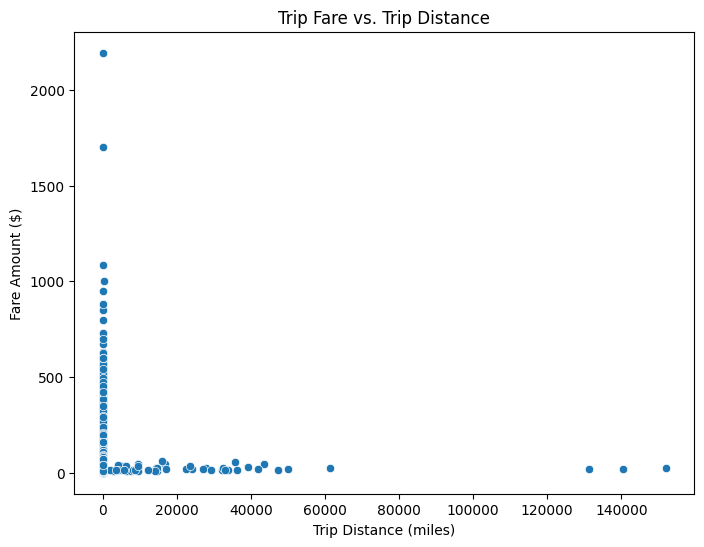
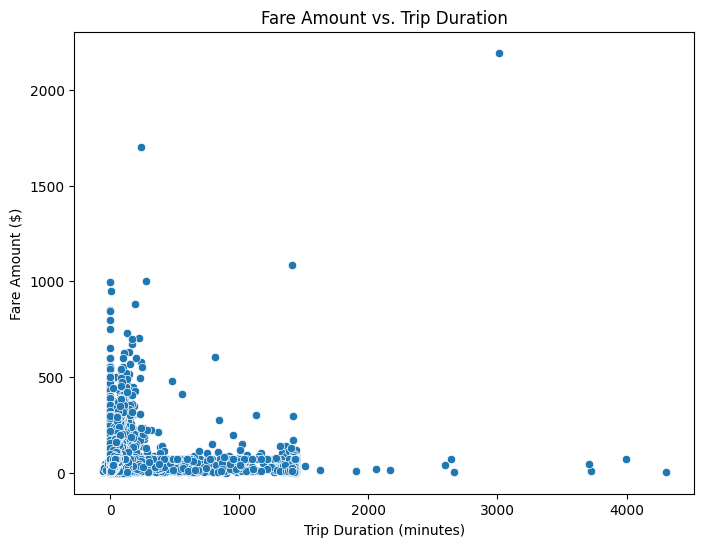
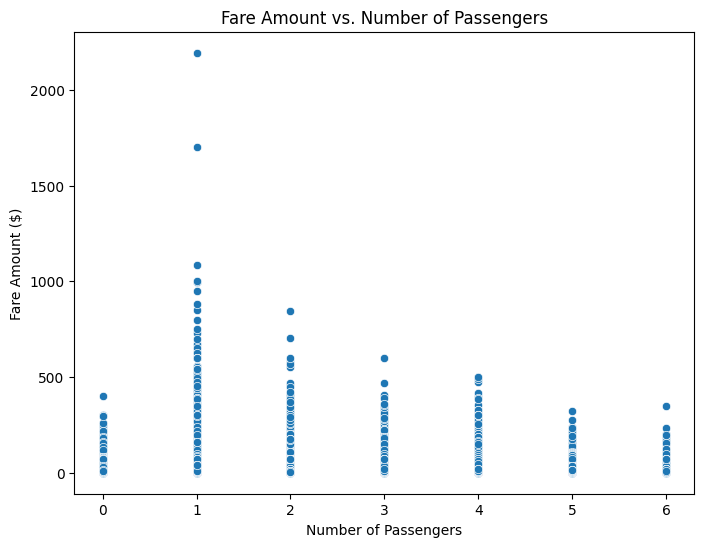
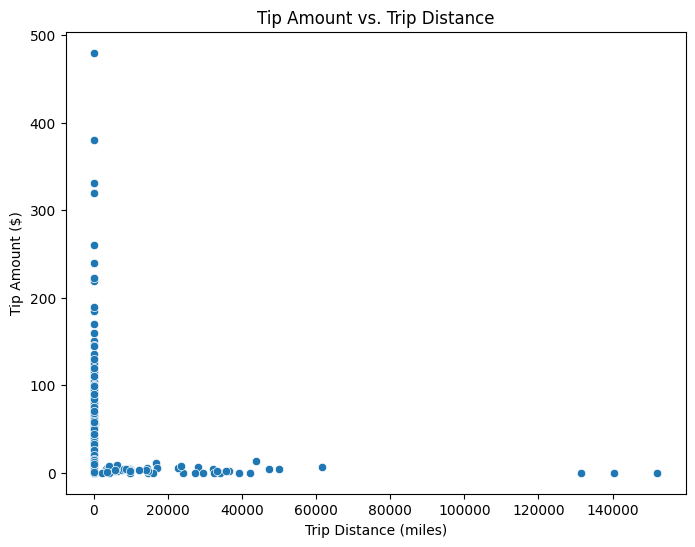
* + 1. **Analyse the monthly revenue trends**
    2. **Find the proportion of each quarter’s revenue in the yearly revenue**The quarterly proportion of revenue are as follow

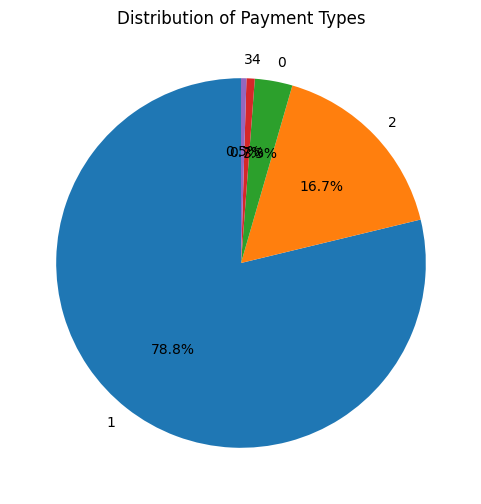
Q1: 0.23

Q2: 0.27

Q3: 0.23

Q4: 0.27

* + 1. **Analyse and visualise the relationship between distance and fare amount**
    2. **Analyse the relationship between fare/tips and trips/passengers**  
    3. **Analyse the distribution of different payment types**



* 1= Credit card
* 2= Cash
* 3= No charge
* 4= Dispute

* + 1. **Load the taxi zones shapefile and display it**
    2. **Merge the zone data with trips data**
    3. **Find the number of trips for each zone/location ID**
    4. **Add the number of trips for each zone to the zones dataframe**
    5. **Plot a map of the zones showing number of trips**
    6. **Conclude with results**

### Detailed EDA: Insights and Strategies

* + 1. **Identify slow routes by comparing average speeds on different routes**

**Below is the data after calculating the slowest route by hour of the day**

Slowest Routes by Hour of the Day:

PULocationID DOLocationID pickup\_hour speed\_mph

0 265 35 0 -42480.0

1 265 144 0 -774.0

2 1 1 0 0.0

3 4 42 0 0.0

4 4 116 0 0.0

.. ... ... ... ...

115 265 9 23 -7068.0

116 1 1 23 0.0

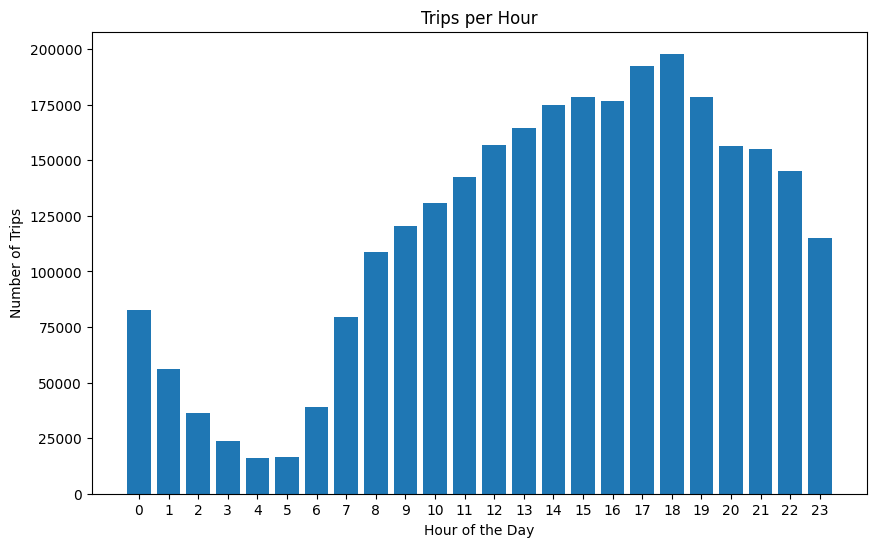
117 4 166 23 0.0

118 7 23 23 0.0

119 7 258 23 0.0

In summary, recognizing high-traffic, high-demand routes empowers taxi companies to make data-driven decisions, optimize their operations, and enhance the overall experience for both drivers and passengers, leading to a more efficient and profitable business.

* + 1. **Calculate the hourly number of trips and identify the busy hours**



Top 5 Busiest Hours:

pickup\_hour

18 197949

17 192256

19 178507

15 178270

16 176804

Name: pickup\_hour, dtype: int64

* + 1. **Scale up the number of trips from above to find the actual number of trips**Estimated Actual Trip Counts for Top 5 Busiest Hours:

pickup\_hour

18 2.639320e+06

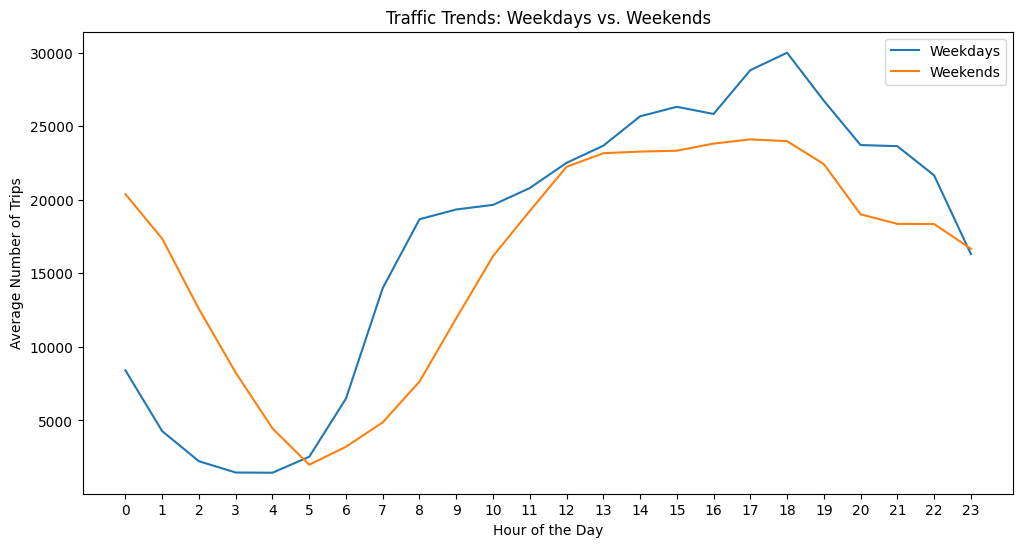
17 2.563413e+06

19 2.380093e+06

15 2.376933e+06

16 2.357387e+06

Name: pickup\_hour, dtype: float64

* + 1. **Compare hourly traffic on weekdays and weekends**

**\*\*Weekday Rush Hours:\*\*** Weekdays show a distinct peak in taxi demand during the morning and evening rush hours. This aligns with typical commute patterns in a major city like NYC, where people rely on taxis to get to and from work or other daily activities.

**\*\*Weekend Evenings: \*\***Weekends exhibit a surge in taxi demand during the evening hours, likely driven by social activities, nightlife, and entertainment events. This suggests that people are more likely to use taxis for leisure and outings during weekends.

**\*\*Weekday Lulls:\*\*** Taxi demand tends to dip during the midday hours on weekdays, indicating a period of lower activity as people are generally at work or school.

**\*\*Weekend Mornings\*\***: Weekends experience a slower start in the morning, with taxi demand picking up gradually as the day progresses. This reflects a more relaxed pace of life on weekends.

looking up for the busy and quiet hour will help to place the cabs at the proper time at proper zone where there is demand and there is less idle hors.

it will also help the taxi companies to use dynamic rate card where they can include the extras as per the demand of the taxies.

the company can predict high demand times and location and deployment more taxies making it easier to hire a taxi making a very perfect customer expeierince by reduing the wait time.

* + 1. **Identify the top 10 zones with high hourly pickups and drops**

Top 10 Pickup Zones:

PULocationID

132 146376

237 133500

161 131270

236 119221

162 100684

186 96724

138 96698

142 93744

230 93686

170 83855

Name: count, dtype: int64

Top 10 Dropoff Zones:

DOLocationID

236 124978

237 119542

161 110834

230 87295

170 84077

162 80782

142 79908

239 78623

141 74563

68 71783

Name: count, dtype: int64

* + 1. **Find the ratio of pickups and dropoffs in each zone**

Top 10 Pickup/Dropoff Ratios:

PULocationID pickup\_dropoff\_ratio

126 132 4.198726

192 199 4.000000

132 138 2.629664

179 186 1.553375

108 114 1.378754

42 43 1.363191

242 249 1.350846

156 162 1.246367

98 100 1.203670

Bottom 10 Pickup/Dropoff Ratios:

PULocationID pickup\_dropoff\_ratio

109 115 0.025000

26 27 0.036364

94 96 0.042553

0 1 .043523

57 58 0.046512

63 64 0.051913

104 109 0.052632

56 57 0.055556

250 257 0.062230

193 200 0.067427

* + 1. **Identify the top zones with high traffic during night hours**

Top 10 Pickup Zones during Night Hours:

PULocationID

132 146376

237 133500

161 131270

236 119221

162 100684

186 96724

138 96698

142 93744

230 93686

170 83855

Name: count, dtype: int64

* + 1. **Find the revenue share for nighttime and daytime hours**

Nighttime Revenue Share: 0.12

Daytime Revenue Share: 0.88

* + 1. **For the different passenger counts, find the average fare per mile per passenger**

verage Fare per Mile per Passenger for Different Passenger Counts:

passenger\_count

0.0 inf

1.0 inf

2.0 inf

3.0 inf

4.0 inf

5.0 inf

6.0 inf

Name: fare\_per\_mile\_per\_passenger, dtype: fl

* + 1. **Find the average fare per mile by hours of the day and by days of the week**

Average Fare per Mile by Day of the Week:

tpep\_pickup\_datetime

Friday inf

Monday inf

Saturday inf

Sunday inf

Thursday inf

Tuesday inf

Wednesday inf

Name: fare\_per\_mile, dtype: float64

* + 1. **Analyse the average fare per mile for the different vendors**

Average Fare per Mile for Different Vendors:

VendorID

1 inf

2 inf

6 6.832511

Name: fare\_per\_mile, dtype: float64

* + 1. **Compare the fare rates of different vendors in a distance-tiered fashion**

**Average Fare per Mile by Vendor and Distance Tier:**

**VendorID distance\_tier fare\_per\_mile**

**0 1 Short 9.940908**

**1 1 Medium 6.387433**

**2 1 Long 4.427908**

**3 2 Short 17.827495**

**4 2 Medium 6.540752**

**5 2 Long 4.492938**

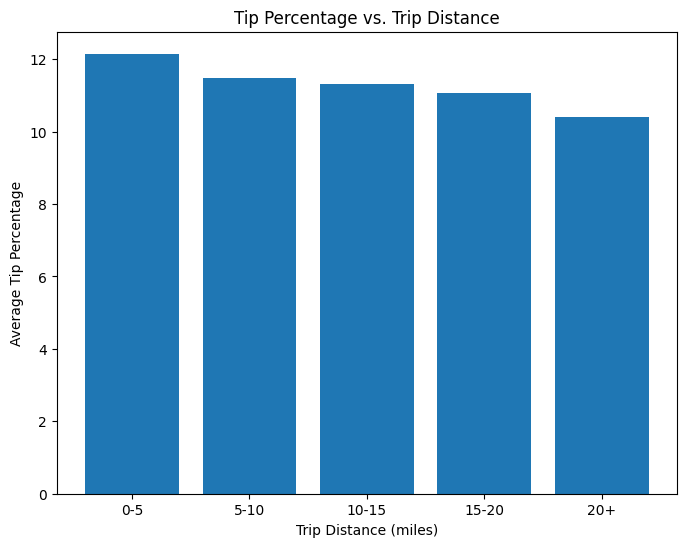
**6 6 Short 32.308362**

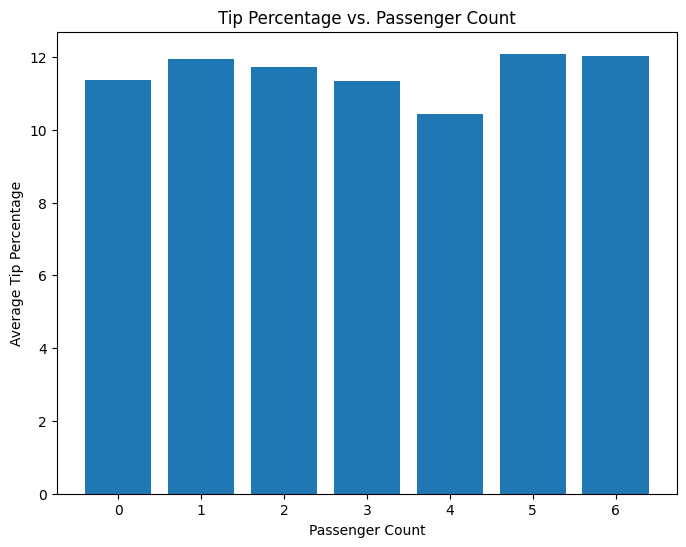
**7 6 Medium 8.308918**

**8 6 Long 4.404028**

Average Fare per Mile for Distances from 2 to 5 miles: 6.503169804972502

Average Fare per Mile for Distances more than 5 miles: 4.477869198214971

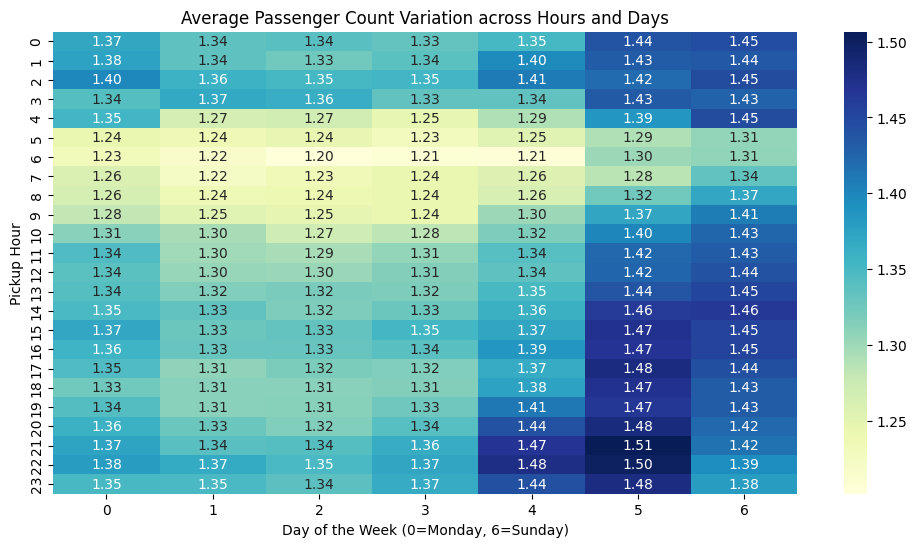
* + 1. **Analyse the tip percentages**



A line graph with numbers and a line

AI-generated content may be incorrect.

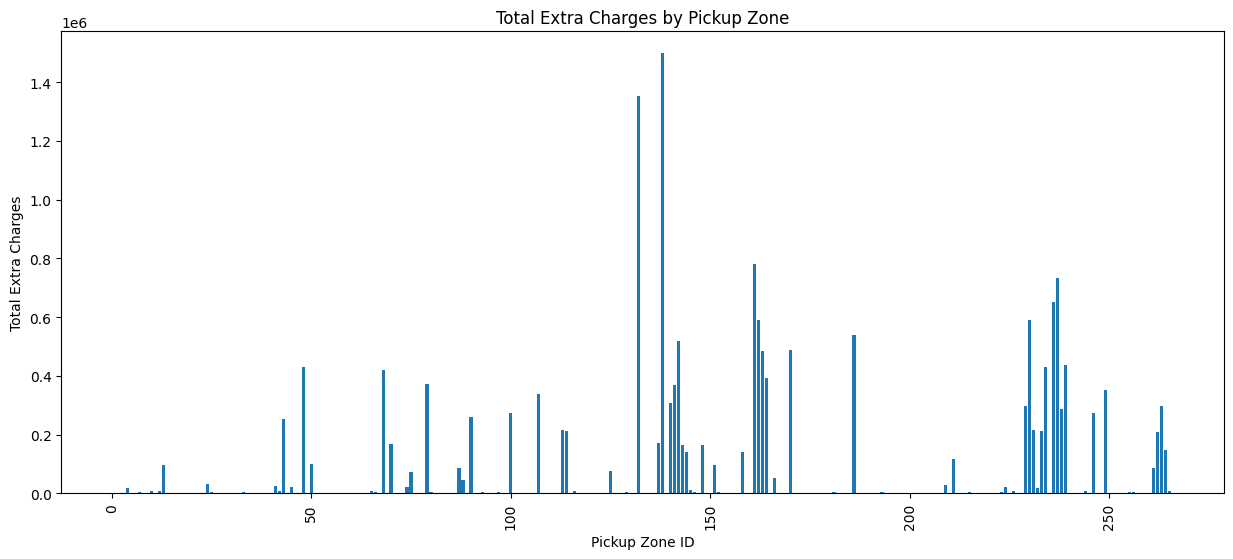
* + 1. **Analyse the trends in passenger count**



* + 1. **Analyse the variation of passenger counts across zones**

A graph of a passenger

AI-generated content may be incorrect.

* + 1. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.**

## Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**

1. Address Operational Inefficiencies:

**\*\*Slow Routes: \*\*** need to identify slow routes and figure a way toavoid these routes.

**\*\*High-Traffic Zones:\*\*** need to optimzie usage of alorithum to makesure the taxis are properly routed through high traffic zone.

Pickup/Dropoff Imbalances: to reduce the Pickup dropoff imbalance by startigically placing the tai in the high pickup zones.

2. Optimize Dispatching Based on Demand:

Peak Hours: Increase the number of taxis available during peak hours to meet the higher demand. This could involve adjusting driver shift schedules or implementing surge pricing to incentivize drivers to work during those times.

Zone-Specific Strategies: Implement zone-specific dispatching strategies based on the identified demand patterns. For example, prioritize dispatching taxis to zones with high pickup demand during specific hours or days.

* + 1. **Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.**

**\*\*Hourly Peaks\*\***: During the busiest hours (16-18hrs), strategically position more cabs in high-demand zones (identified in 3.2.5). These could include popular pickup locations like airports, business districts, and entertainment areas.

**\*\*Hourly Lulls\*\***: During off-peak hours ( 4-5hrs), reduce the number of cabs in less active zones to minimize idle time. Encourage drivers to reposition to areas with anticipated demand increases based on upcoming events or predictable shift changes.

**\*\*Day-of-Week Variations\*\***: Increase cab density in areas with high demand on specific days. For example, focus on business districts during weekdays and entertainment hubs during weekends.

**\*\*Monthly Trends:\*\*** Analyze monthly trends to anticipate fluctuations in demand. During peak months, ensure higher cab availability in popular areas. During slower months, consider strategies to incentivize drivers to remain active or adjust fleet size accordingly.

* + 1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**

1. Dynamic Pricing:

Demand-Based Adjustments: Implement dynamic pricing models that adjust fares based on real-time demand and supply.

2. Competitive Pricing Analysis:

Monitor Competitors: Continuously monitor the pricing strategies of other transportation vendors, such as ride-hailing services, to ensure that taxi fares remain competitive.

Route Optimization: Encourage drivers to utilize navigation apps or dispatching software that provides optimal routes to minimize travel times and fuel costs. This can help reduce operating expenses and potentially allow for lower fares while maintaining profitability.

Idle Time Reduction: Implement strategies to reduce driver idle time, such as repositioning taxis to high-demand zones during off-peak hours or offering incentives for completing trips to areas with low taxi availability. This can improve driver utilization and potentially allow for lower fares without impacting driver earnings.

4. Customer Segmentation and Promotions:

Loyalty Programs: Implement loyalty programs to reward frequent taxi riders with discounts or exclusive benefits. This can help retain existing customers and incentivize repeat business.

Targeted Promotions: Offer targeted promotions or discounts for specific trip types or passenger segments, such as airport trips, business travelers, or students. This can help attract new customers and increase ridership during specific times or in specific zones.

Bundled Services: Consider offering bundled services, such as pre-booked airport transfers or corporate accounts, at discounted rates to incentivize larger volume bookings.

5. Data-Driven Evaluation:

Monitor Key Metrics: Continuously monitor key metrics, such as average fare, trip volume, passenger wait times, and driver earnings, to assess the impact of pricing adjustments. This data-driven approach allows for ongoing evaluation and fine-tuning of the pricing strategy to maximize revenue and maintain competitiveness.

A/B Testing: Conduct A/B testing to compare the effectiveness of different pricing strategies in specific zones or time periods. This can help identify the optimal pricing approach for different market segments and demand patterns.

Feedback Mechanisms: Establish feedback mechanisms for passengers and drivers to gather insights and understand their perspectives on pricing and service quality. This valuable feedback can inform future pricing adjustments and service improvements.

Key Considerations:

Transparency: Maintain transparency with passengers regarding pricing adjustments and surge pricing to avoid negative perceptions or surprises. Clearly communicate the factors that influence fare calculations.

Regulatory Compliance: Ensure that all pricing strategies comply with relevant regulations and guidelines set by the NYC Taxi and Limousine Commission (TLC).

Driver Collaboration: Involve drivers in the pricing strategy discussions and consider their feedback to ensure that fare adjustments are fair and incentivize them to provide quality service.

By implementing these data-driven pricing adjustments, taxi companies can optimize their fares to maximize revenue while maintaining competitive rates with other transportation vendors. This strategic approach can contribute to the long-term success and sustainability of the taxi industry in NYC.